

KONRAD-ADENAUER-STIFTUNG · MEDIA PROGRAMME SUB-SAHARAN AFRICA

Reporting on AI-driven disinformation

A working guide for journalists

Heather A. Thuynsma, Yanga Malotana
and Christopher Nehring



Reporting on AI-driven disinformation

A working guide for journalists



SECTIONS	REFERENCE CARDS	VERIFICATION STEPS	SECONDS TO ARCHIVE
8	4	7	30

Konrad-Adenauer-Stiftung,

Media Programme Sub-Saharan Africa

60 Hume Road, Dunkeld 2196,

Johannesburg,

Republic of South Africa

Telephone: + 27 (0)11 214-2900

www.kas.de/mediaafrica

Publication © Konrad-Adenauer-Stiftung e.V., 2026

Text © Heather A. Thuymsma, Yanga Malotana and

Christopher Nehring 2026

All rights reserved.

Requests for review copies and other enquiries concerning this publication are to be sent to the publisher.

The responsibility for facts, opinions and cross-references to external sources in this publication rests exclusively with the contributors and their interpretations do not necessarily reflect the views or policies of the Konrad-Adenauer-Stiftung.

Cover design: Stephen Symons

Typography and design: Stephen Symons

ISBN: 978-1-0671823-4-2 (Print)

ISBN: 978-1-0671823-5-9 (Digital)

Reporting on AI-driven disinformation

A working guide for journalists

HOW TO USE THIS GUIDE

We built this guide to be opened at the point of need, not read front to back. Each section answers one question and closes with a checklist you can apply in a few minutes.

IF YOU ARE ASKING ...	GO TO
Is this content actually manipulated, or am I over-reading it?	01 — Spot it
How do I check whether this video, image or audio is real?	02 — Check it
How do I make sure this evidence still exists tomorrow?	03 — Preserve it
Should I report this at all, or would reporting it spread it?	04 — Decide
How do I write this up without amplifying the false claim?	05 — Write it
Why does my correction never reach the people who saw the lie?	06 — Distribute it
Why does this beat leave me feeling like this?	07 — Protect yourself
I need a one-page checklist right now.	08 — Reference cards
I work in campaign communications, not journalism.	Appendix A
I want the background on how all this works.	Appendix B

IF YOU HAVE FIVE MINUTES AND A DEADLINE

- 1 **Archive the content before you do anything else** (section 03). Thirty seconds. If it disappears while you are deciding, you have lost the story.
- 2 **Trace it back to whoever posted it first.** The account you saw it on is almost never the source.
- 3 **Ask what you can establish and write only that.** “We could not verify” is publishable. “Reportedly” is not verification.
- 4 **If you publish, lead with what is true,** not with the false claim.

SECTION ONE

01

Spot it: is this manipulation?

Manipulation is the deliberate engineering of a false impression, and it can happen either in the content or in its spread. Three questions establish what kind of problem you are looking at, because the answers decides whom you hold responsible.

THE THREE-QUESTION TEST

Q1 Is the content false?

If it is true but deployed to cause harm (a real document leaked to damage someone, a genuine recording released at a calculated moment), then you are handling the harmful use of true material. Your test is then not “is this accurate?” but “does publishing it serve the public, or mainly deliver the harm the leaker intended?”

Q2 If it is false, was it meant to deceive?

If it was, someone made a deliberate choice, and that person is the story. If it was not, and the people passing it on believe it, then the story is how the information spread and what needs to be corrected, rather than the sharer.

Q3 Who made it, and who pushed it?

These are usually different people, and the second person is often the more newsworthy. A fabricated video matters far less than the account with two million followers that chose to post it.

NAMING IT ACCURATELY IS NOT TAKING SIDES

Accuracy is not neutrality. Naming disinformation as disinformation is not a political act. It is the primary professional obligation of a journalist. False balance between a verified fact and a deliberate fabrication does not serve the public; it serves the fabricator. And writing “a claim made by some” about content you have established was deliberately fabricated is not caution. It is a decision to protect whoever made it.

THE THREE QUESTIONS, AND WHAT EACH ONE DECIDES

Q1

IS IT FALSE?

If it is true but weaponised, you are handling the harmful use of true material rather than a fake.

Q2

MEANT TO DECEIVE?

If yes, the person who chose to make it is the story. If no, the spread and the correction are.

Q3

WHO PUSHED IT?

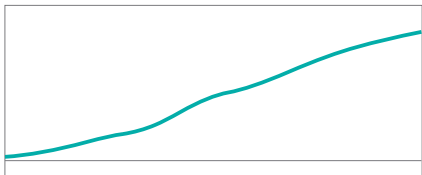
Usually not the maker — and usually the more newsworthy of the two.

OUTPUT Whom you hold responsible, which is what the story is actually about.

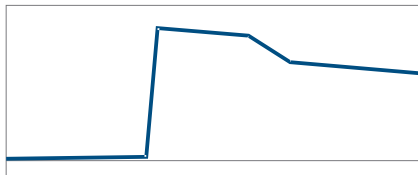
SIGNS THAT A TREND IS BEING MANUFACTURED

A manufactured trend is one whose apparent public support is engineered rather than earned. Volume is not evidence of public interest. A message can be entirely legitimate on its own terms and still become an act of manipulation when 50,000 automated accounts push it. What you are looking for is the shape of the spread, not the wording of the post.

THE SHAPE OF THE SPREAD



ORGANIC, BUILDING SLOWLY AND UNEVENLY



MANUFACTURED — A VERTICAL SPIKE FROM NOTHING

POSTING TIMESTAMPS, FIRST TWENTY ACCOUNTS

ORGANIC



COORDINATED



A coordinated push leaves a signature in the posting times. Clusters of accounts that post the same material within seconds of one another often do so at intervals that no human rhythm produces. Export or hand-record the posting times for the first twenty accounts, and the pattern usually surfaces without any specialist tooling.

WHAT TO CHECK

WHAT A MANUFACTURED TREND LOOKS LIKE

Timing

A steep vertical spike from nothing. Organic interest builds more slowly and unevenly. Check whether the first wave of posts appeared within a few minutes of each other.

The accounts

Recently created, posting constantly, following far more accounts than follow them, little history unrelated to this topic.

The wording

Identical or near-identical text across accounts that claim no connection. Near-identical wording with small variations also matters because AI-generated copy produces slight changes in text where copy-and-paste would produce exact matches.

Hashtags

A brand-new hashtag appearing fully formed across dozens of accounts, or an existing hashtag suddenly carrying content unrelated to its original use.

Images

The same image or video, re-posted with slight crops or filters to defeat automated matching.

Emotional charge	High outrage, low specificity. Content engineered for sharing rather than for informing.
The platform	On WhatsApp you can't see engagement signals. Assess spread inside closed groups by asking people what they are receiving.

If three or more of these are present, treat coordination as your working assumption and report it as part of the story. And treat the platform that carried it as one of the actors. A fabricated video that reached 3 million people because a recommendation system rewarded its outrage value is a story about the platform as much as about whoever made it.

SECTION TWO

02 Check it: verification

Verification is the establishment of a claim through independent sources that do not share an origin. That core rule has not changed; two outlets quoting the same press release remain one source. What has changed is the evidentiary status of the recording itself. You can no longer assume that a video of a person saying something is evidence that the person said it.

THE SEVEN STEPS

01 Archive at first contact

The moment content becomes potentially relevant, preserve it. Do not wait until you have decided to publish. Save the URL, screenshot the post including the URL, timestamp and account details, and record the follower count and engagement figures visible at that moment.

02 Find the original

Trace the chain of shares backwards. Establish when it was first posted, by whom and on what platform. The account you encountered it on is rarely the source.

03 Triangulate

Two independent primary sources with no shared origin. If you can't find those, then the claim is alleged rather than verified, and you must write it that way.

04 Run the synthetic content checks

Run reverse image searches on video frames and images, and extract the metadata. Detection tools should be treated as indicative only.

05 Verify the person independently

If the content claims to show a named individual, confirm through that person's office, their verified accounts, or someone who was present. Never from the content itself.

06 Apply the reverse test

If a political figure claims that genuine content is "AI-generated" without producing forensic evidence, that claim is the story. Apply the same standard to an unevidenced fabrication claim that you apply to an unevidenced factual claim.

07 Write down what you could and could not establish

Do this before filing. If you can't verify it, either leave it unpublished or publish the fact that it is circulating and unverifiable. The story is then about the circulation.



TOOLS

None of these tools gives you an answer. They give you evidence, which you must then weigh. The single most common error in AI-era reporting is treating a detector's percentage as a finding.

TRACING AND MATCHING	WHAT IT DOES	NOTES
Google Images, TinEye, Yandex	Reverse image search	Run all three. They index differently, and Yandex is frequently better on faces
InVID / WeVerify	Breaks video into keyframes for reverse search; metadata and magnifier tools	This is the standard first move on any video
ExifTool	Extracts embedded metadata: camera, timestamps, GPS, editing software	Most platforms strip metadata on upload, so absence proves nothing. Presence, on the other hand, can be decisive
FotoForensics	Error-level analysis on images	Shows inconsistent compression that can indicate editing. Easy to misread. Treat it as a prompt to look harder, not as proof
AI DETECTION	COVERAGE	ACCESS
DeepFake-o-Meter (University at Buffalo)	Image, video and audio. Runs several published research detectors at once and shows you where they disagree	Free, open-source, academic. Start here, because the disagreement between detectors is itself informative

Hive Moderation	Image, video, audio, AI-text	Commercial, free demo tier
AI or Not	Image, audio	Commercial, limited free checks
Sightengine	Image, video	Commercial, API-first
Reality Defender	Image, video, audio, text	Commercial, enterprise pricing; used by several newsroom partnerships

HOW TO WRITE UP A DETECTOR RESULT

DO NOT WRITE “Analysis confirmed the video was AI-generated.”

WRITE “Three detection tools returned results consistent with AI generation. Detection tools are not conclusive and can be wrong in both directions. The stronger evidence in this case was the original 2017 footage / the person’s office confirming they were elsewhere / the metadata.”

Detection performance drops sharply on compressed, re-uploaded and non-English content, which describes almost everything that circulates. Your provenance work is more reliable than any detector, and it is what you should lead with.

A CAUTION ON AFRICAN-LANGUAGE MATERIAL

Newsroom research documents that AI tools used in African newsrooms regularly fail on context, nuance and political terminology in African languages. This is because indigenous African-language models and detection tools do not exist. In fact, detection capacity lags behind production capacity by a wider margin than it does in English-language environments (Allen and Nehring, 2025). Do not rely on an AI-assisted check for content in a language the tool handles poorly. Use a human speaker with local political knowledge.

SECTION THREE

03

Preserve it: archiving before it disappears

Archiving is the act of fixing a piece of evidence in the state in which you found it. It is the cheapest insurance available, and journalists skip it more often than any other step.

People believe that the internet forgets and that content, once deleted, is gone. But that is not quite right. Deleted material is very often recoverable because archives, caches, re-posts and specialist tools mean that something once public can usually be found again. Recovery is not guaranteed, however: it consumes time you may not have on deadline, and it may not capture the version you actually saw. So archive immediately, because thirty seconds now saves you two hours later and preserves the state you are reporting on.

WHAT TO DO AT FIRST CONTACT

01

Save the page to the Wayback Machine (Save Page Now at archive.org) and to archive.today. Use both. They behave differently, and archive.today captures some dynamic pages the Wayback Machine misses.

02

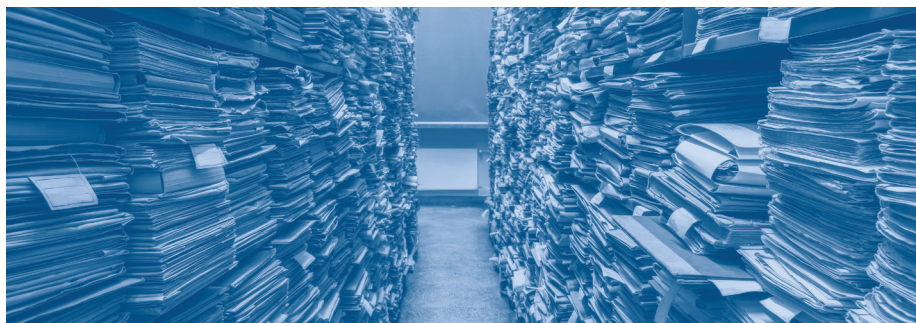
Screenshot the full window including the URL bar, the system clock, the account name and verification badge, and the engagement count.

03

Save the media file itself, not a screenshot of it. A screenshot of a video destroys the metadata and the compression signature you may later need.

04

Note the time you captured it in your story file.



RECOVERING MATERIAL THAT IS ALREADY GONE

TOOL	USE IT FOR
Wayback Machine	Historical versions of web pages. The CDX API lets you list every capture of a URL with dates, which is how you find the version that existed on a given day.
archive.today	Pages the Wayback Machine did not capture, and dynamic pages. Often the only surviving copy of a deleted article.
CachedView	Queries several caches and archives at once. A fast first sweep when you do not know which archive holds it.
Hunchly	Browser-based capture that records everything you visit during an investigation with timestamps and hashes, building an audit trail. Paid. Worth it for sustained investigations.
Bellingcat Auto Archiver	Bulk automated archiving of social media links, for when you are tracking a campaign rather than a single post.
Platform-native search	Quote-posts, screenshots and angry replies frequently outlive the deleted original. Search the phrasing rather than the URL.

THE DELETION IS THE STORY

When a political actor deletes content after publication, the deletion itself is newsworthy: the original wording, the fact of removal and the timing relative to whatever prompted it. But you can report this only if you archived first. An unexplained deletion, published alongside the archived original, is one of the most straightforward accountability stories available to a journalist.

SECTION FOUR

04

Decide: should you report this at all?

Reporting on a false claim spreads it. No framing is clever enough to escape that fully, which makes publication itself the first editorial decision rather than a formality that follows verification. And the decision needs criteria that can be acted on under a deadline.

WHEN TO REPORT

- **It has already reached your audience.** If your readers have seen it, silence does not protect them. It leaves the false version as the only version they hold.
- **You can name someone accountable.** A story that identifies who made or pushed the content exposes more than a generic “false content is circulating” story.
- **There is a demonstrated harm.** People acted on it, an institution responded to it, an outcome changed.
- **The technique is new or newly effective,** and explaining how it works will help your audience recognise the next one.

WHEN NOT TO REPORT

- **It is contained.** If it is circulating in a small, closed network and your audience has not encountered it, then publishing is the act that introduces it to them. This is the most common unforced error journalists make.
- **You can’t attribute it.** Ask what the story is beyond a restatement of the false claim.
- **The claim is more memorable than your correction.** Some fabrications are vivid enough that the correction functions as a delivery mechanism.
- **You would be the amplification event.** If your outlet is substantially larger than the reach the content achieved on its own, you are not covering the spread. You are the spread.
- **It is one of many.** High-volume operations flood the zone with contradictory fabrications so that each one gets individually chased. Cover the operation.

VOLUME IS A TACTIC IN ITSELF

- Disinformation does not only win by being more compelling than the truth. It also wins by exhausting you, running ten fake narratives about everything and everyone, so that no single one can be run down and the reader concludes that nothing can be known.
- It wins by repetition. A claim repeated often enough starts to feel familiar, and readers routinely mistake familiarity for truth. This is why the same fabrication returns in the same words months later. It is not laziness. It is the mechanism.

Both have the same practical consequence. The unit of coverage should usually be the campaign, not the individual claim.

SECTION FIVE

05 Write it: the truth sandwich

Structure is an editorial decision about what your reader retains. Once you have decided to publish, the order in which you present the truth and the falsehood determines whether the reader leaves with the accurate account or with the fabrication.

THE STRUCTURE

- 01** Lead with the accurate framing

State what is true first and prominently.
- 02** Identify the false claim briefly, with explicit attribution

“Party X claims...”, with named actors, kept short and never as a standalone quotable sentence.
- 03** Return to the accurate framing

Reinforce what is true at the close. The last line is the one a reader will retain.

THE HEADLINE IS WHERE MOST DEBUNKS FAIL

HEADLINE	WHAT IT DOES
“Trump endorses MK”	Repeats the fabrication as though it were the news. Even with a correction below, this is the sentence that gets screenshotted and shared.
“Viral video claims Trump endorses MK”	Better, but attributes the claim to a faceless crowd. Nobody is accountable, and “viral” implies organic public interest that may not exist.
“No, Donald Trump has not backed South Africa’s MK party”	Leads with the truth, in a form that survives being read alone. This is the structure Africa Check uses.

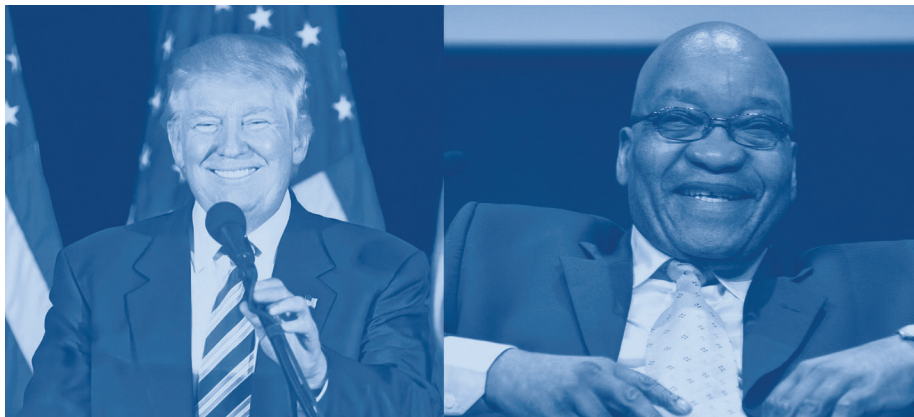


IMAGE SOURCE: THE AFRICA REPORT

WORKED EXAMPLE

What follows is a model of a finished piece, reconstructed from the documented Africa Check and ISS/Murmur Intelligence material on the 2024 South African case (Allen and le Roux, 2024). We have written it as a model rather than reproducing it verbatim. The turquoise notes are annotation, not copy.

MODEL COPY

No, Donald Trump has not endorsed the MK party

The video circulating on X and TikTok uses footage from a 2017 NBC interview, with audio generated by AI

Former US president Donald Trump has made no statement about South Africa's uMkhonto weSizwe party, and has not endorsed any party contesting the 2024 election.

A video shared this week appears to show him doing so. It surfaced first on TikTok and then spread on X, where the MK party figure Duduzile Zuma-Sambudla posted it on 9 March, nearly three months before polling day. It has been viewed tens of thousands of times across TikTok, X, Instagram and YouTube.

The video is fabricated. The visual footage comes from an interview Trump gave to NBC's Lester Holt at the White House in 2017, identifiable by the presidential flag behind him, the angle of his face, his tie and the flag pin on his jacket, and by a reverse

TRUTH FIRST. THE READER WHO READS ONLY THIS LINE LEAVES WITH THE ACCURATE POSITION

THE CLAIM. BRIEF, ATTRIBUTED TO NAMED ACTORS, AND LOCATED IN A SPECIFIC COMMUNITY RATHER THAN DESCRIBED AS "VIRAL"

THE EVIDENCE. HOW WE KNOW, IN TERMS THE READER CAN CHECK INDEPENDENTLY

image search on individual frames that returns the original broadcast. The audio does not match that footage. The lip movements do not correspond to the words, and the speech carries the flat intonation and clipped word endings characteristic of AI voice synthesis. And the endorsement is an anachronism: the MK party was not registered until September 2023, six years after the interview was filmed.

The technique is not sophisticated. Anyone with a consumer voice-cloning tool and a few minutes of public speech can produce a comparable file. What made this one effective was not the quality of the fake but the speed and reach of its distribution within a community already receptive to the message. And the wider pattern matters more than this clip. The most systematic study of the 2024 campaign found the use of generative AI and deepfakes limited, while documenting extensive manipulation of distribution through coordinated amplification, paid and undisclosed influencers, and sock-puppet accounts (Allen and le Roux, 2024).

THE MECHANISM. THE PART THAT PROTECTS YOUR READER AGAINST THE NEXT ONE, AND THE PART THAT POINTS AT WHAT THE CAMPAIGN WAS ACTUALLY DOING

Two things to check when you see a clip like this: run a reverse image search on a still frame, which takes under a minute and would have identified the 2017 source; and check whether any established news organisation has reported the statement, because an endorsement by a former US president would not appear first as an unsourced video on TikTok.

WHAT THE READER CAN DO

Trump has not endorsed the MK party, and nobody has produced evidence that he has commented on the South African election at all.

TRUTH LAST

Note what the model does not do. It never states the false claim as a standalone sentence that could be quoted on its own. It names the amplifier rather than describing the content as “widely circulating”. It explains the method, not just the verdict. And it closes on the truth.

BEFORE YOU FILE

ACCURACY	Two independent primary sources for every factual claim; anything unverified explicitly labelled.	ATTRIBUTION	Named actors, not “some users”. A story that names nobody protects everybody.
FRAMING	Truth first, claim second and attributed, truth last. Check the headline separately because it travels on its own.	AMPLIFICATION	Will this reach people who have not seen the claim, and is that justified?
DISTRIBUTION	How will this reach the people who did see it? (Section 06.)	HARM	What damage has been done, and could your piece add to it?
PRESSURE	Has anyone leaned on you to soften or drop this? Document it.	AI TOOLS	If you used them to research, translate or draft, have you checked their output for absorbed framing and invented sources?

SECTION SIX

06

Distribute it: why corrections do not reach the people who need them

Distribution is the point at which most careful debunking work is wasted. It is also the part of the process that newsrooms treat as somebody else's job.

The Institute for Security Studies (ISS) analysis of the 2024 South African election describes the mechanism. Platform ranking promotes content from accounts and groups a user already engages with, reinforcing information silos in which, as the report puts it, users “are exposed to content that aligns with their existing beliefs, making it difficult for fact-checked or corrective information to penetrate these networks” (Allen and le Roux, 2024). The same report notes that even after a claim is debunked, it tends to keep circulating among a party's own supporters.

So the problem is not that your correction is wrong or late. It is that the network that carried the false claim is structurally disinclined to carry your correction. This is not an argument against publishing corrections. The same report credits fact-checking with mitigating some campaigns, and finds professional media to be a significant bulwark. It is an argument that a correction without a distribution plan is an unfinished piece of work. The people running the original campaign thought hard about distribution. If you do not, you are competing on unequal terms and losing for reasons that have nothing to do with the quality of your reporting.

PLAN THE DISTRIBUTION BEFORE YOU WRITE

Establish where the claim actually spread: which platform, which communities, which language, which closed groups. This is reporting work, and it changes what you write, not merely where you put it.

Publish in the format that platform carries

A 1,200-word web article does not travel in a WhatsApp group. A short voice note, a single image with the correction legible without tapping through, or a forwarded text block does. Produce that version yourself rather than hoping someone makes it.

Do not depend on links

In closed-messaging environments a link is friction, and recipients often distrust it. The correction has to be intelligible from the message itself.

Use trusted messengers

A correction carried by someone the community already trusts (a local radio presenter, a community organisation, a religious or civic leader) outperforms the same words from an outlet the community regards as hostile. Build those relationships before you need them.

Translate

If the claim spread in isiZulu or Hausa, an English correction is not a correction for that audience. Have the translation checked by a speaker; machine translation of political terminology is where accuracy quietly disappears.

Republish on the anniversary

Fabrications return, usually in the same words. A correction that is already indexed and dated when the claim resurfaces is more important than a new one written under pressure.

WHY THE CORRECTION DOES NOT TRAVEL THE SAME ROAD

THE NETWORK THAT CARRIED THE CLAIM

Accounts and groups the user already engages with. Ranking keeps feeding them more of the same.


BLOCKED BY
PLATFORM RANKING

YOUR CORRECTION

Accurate, sourced, published, and structurally unlikely to enter that network at all.

Schematic. The fix is not a better correction. It is a distribution plan that does not depend on that road.

BEING FINDABLE, BY PEOPLE AND BY MACHINES

An increasing share of what audiences encounter is mediated by search ranking and by AI assistants summarising the web. Both retrieve from what is publicly indexed. If fabricated content is well-structured and abundant, and your correction is neither, the summary your audience receives will reflect the fabrication.

State the correction as a plain declarative sentence early in the piece ("Trump did not endorse the MK party"), because that is the sentence a retrieval system can lift. Rhetorical or clever openings are invisible to this process.

Include the false claim's wording once, clearly labelled as false, so that a search for those words reaches your correction. Omit it entirely and the people searching for the claim never find you.

Use clear headings, dates and named entities. Structure is what makes a page retrievable.

Keep the piece accessible. Corrections behind hard paywalls are absent from the environment in which the false claim is competing, and the fabrication is always free.

THE HONEST LIMIT

None of this equalises the contest. A newsroom producing a verified correction can't match the output of an operation generating content at machine speed across a hundred domains. The realistic goal is to be present and findable in the same space, not to out-produce it. And the structural problem belongs in public and regulatory view rather than being absorbed silently as a professional failing.

SECTION SEVEN

07

Protect yourself

Information-disorder burnout is a distinct occupational condition, and it presents differently from general journalism burnout. The difference matters because the standard remedies (a holiday, a lighter week, a change of assignment) address exhaustion rather than the erosion of epistemic confidence that characterises this beat. Practitioners and the editors who manage them need to know the signs.

FIVE SIGNALS

01 Verification confidence decline

Difficulty trusting your own verification conclusions, second-guessing findings you have established, experiencing verification as unresolvable rather than difficult.

02 Narrative disconnection

The sense that the work is being produced but not received, that corrections are not reaching the people who need them.

03 Bad-faith fatigue

A generalised scepticism that no longer discriminates between genuine and manufactured doubt.

04 Loss of footing

Moments of real uncertainty about your own capacity to evaluate information. Am I becoming unable to tell what is true?

05 Community withdrawal

Retreating from professional community rather than seeking support within it.



WHAT HELPS

Never verify alone when collaboration is possible

The most effective protection is a professional community of practice: colleagues who apply the same standards and can confirm or challenge your conclusions.

Set boundaries on bad-faith engagement

Not every actor who engages with your reporting deserves substantive engagement in return. In other words, engagement whose purpose is to exhaust and discredit.

Distinguish difficulty from erosion

Uncertainty you resolve through further verification is the job working correctly. Uncertainty that persists after you have done the work, and attaches to your own judgement rather than to the evidence, is the signal to attend to.

Keep the methodology record

A clear written record of your verification steps is both professional protection and a way of demonstrating to yourself, on a bad day, that your judgement is sound.

WHAT YOUR NEWSROOM OWES YOU

75%

OF PRACTISING SOUTH
AFRICAN JOURNALISTS
HAVE RECEIVED
NO FORMAL AI
TRAINING FROM THEIR
ORGANISATION
(Allen, Wasserman
and Mbekela, 2026)

Training. This is not an optional extra. They are the conditions that make the rest of this guide useable and it is reasonable to ask for them in writing.

Editorial time for verification. The workflow in Section 02 takes time that commercially pressured newsrooms are structurally reluctant to allocate.

Legal support as standard.

Explicit recognition. In editorial conversation and in performance review, that journalism carries a psychological load.

SECTION EIGHT

08

Quick reference cards

Print these and keep them beside you.

Card 1 — First contact

ACTION

- ☐ Archive to the Wayback Machine and archive.today
- ☐ Screenshot with URL, clock, account, engagement count
- ☐ Save the media file itself, not a screenshot of it
- ☐ Record the platform, account, follower count, posting time
- ☐ Note the time of your capture in the story file

Card 2 — Is it coordinated?

CHECK

- ☐ Vertical spike from nothing rather than a build
- ☐ New accounts, high posting frequency, skewed follower ratios
- ☐ Identical or near-identical wording across unconnected accounts
- ☐ Posting timestamps clustered within seconds
- ☐ Brand-new hashtag appearing fully formed, or a hijacked one
- ☐ Same image reposted with small crops to defeat matching
- ☐ High outrage, low specificity

THREE OR MORE Treat coordination as your working assumption and report it.

Card 3 — Before you file

- ☐ Two independent primary sources for every factual claim
- ☐ Named actors, not “some users”
- ☐ Truth first, claim attributed, truth last
- ☐ Headline works when read alone and does not repeat the fabrication
- ☐ Amplification weighed: will this reach people who have not seen it?
- ☐ Distribution planned: how does this reach people who did?
- ☐ Any pressure to soften or drop the story documented
- ☐ AI-assisted research, translation or drafting checked for absorbed framing and invented sources

Card 4 — Ethical anchors

Accuracy is not neutrality. False balance between a verified fact and a deliberate fabrication does not serve the public; it serves the fabricator.

Speed is not an excuse for publishing unverified content. Correcting a false report does not undo the harm the report has already done.

Republishing amplified content without asking why it was amplified makes the journalist a node in the manipulation network.

The journalist's first obligation is to the public's right to accurate information. Not to access, not to speed, not to engagement metrics, not to institutional relationships.

A practitioner who is depleted can't serve the public interest with the rigour it requires. Self-care in this context is professional maintenance.

APPENDIX

A For campaign and party communicators

The taxonomy of information disorder is not only a diagnostic tool for scrutinising others. It is an audit instrument for your own output.

AUDIT YOUR OWN OUTPUT

- Is the content factually accurate and verifiable? If it is not, and you share it as though it were, your organisation is a source of misinformation regardless of intent.
- Are you framing accurate content selectively to create a misleading impression? That is deception by omission, the letter of accuracy without its spirit.
- Are you deploying true information mainly to damage an opponent rather than to inform? Strategic truth is also an ethical problem.

Here is a practical institutional test. Would your communications team be comfortable if someone published the full production process behind this content (the sourcing, the framing and the distribution decisions) as a news story? If not, that discomfort is diagnostic.

WHERE PERSUASION ENDS

Persuasion is the attempt to move an audience by giving it reasons. Manipulation is the attempt to move an audience by giving it a false picture of the situation it is reasoning about. Strategic communication is, therefore, not the same thing as linguistic manipulation. Framing is legitimate. Emphasis is legitimate. Choosing the most persuasive accurate language is a craft, not a deception. The line runs between language that describes the factual situation accurately and language designed to create a false impression of it.

There is also a strategic argument for accurate language. An electorate with a high baseline of factual understanding is a harder target for your opponents' disinformation campaigns. If you win by manipulating the information environment, then you will govern in the environment you created.

AMPLIFICATION

Amplification is not a neutral distribution choice made downstream of the message. If coordinated or automated inauthentic accounts push a claim that is accurate on its face, the operation is manipulation whatever the claim says. Documenting these networks makes the point very clearly. The technique sits in the distribution layer, not in the wording (Wack, Linvill and Warren, 2024). The legitimacy of your communication does not rest on the content alone. It rests on whether the appearance of public support behind it is real.

PRACTICAL OBLIGATIONS

Assume everything is archived

The assumption that deleting a post removes it from the record is almost always wrong. Plan your communications accordingly.

Correct rather than delete

Deleting a published statement without acknowledgement is a form of record manipulation. Publish the original, the correction, and the reason.

Never use synthetic media without disclosure

Using AI-generated or AI-edited video, audio or images without telling your audience is not a presentational choice; it is a provenance failure. Keep provenance records: what was generated, when, with what inputs, and which human editorial decisions were made about it.

“I was told to” is not a defence

Your professional responsibility is personal. Some requests can't be navigated through careful framing. They are requests to produce disinformation.

APPENDIX

Background: how this works

A practitioner needs the tools first and the theory available afterwards, so this is the compressed version. The full argument, with its evidence and sources, is in the companion academic volume.

THREE KINDS OF INFORMATION PROBLEM

<u>Misinformation</u>	<u>Disinformation</u>	<u>Malinformation</u>
False, and shared by people who believe it.	False, and shared deliberately to deceive.	True, and deployed to cause harm.

The distinction is important because it determines who is accountable, and therefore what your story is about. Misinformation is a problem of correction. Disinformation is a problem of attribution. Malinformation is a problem of editorial judgement, because the material is real and the question is whether publishing it serves anyone but the person who released it.

WHAT AI ACTUALLY CHANGED

The common assumption is that AI made fakes more convincing. That is the smaller half of the change. The larger half is that it made production and distribution cheap enough to industrialise, and cost is what governs how much of this the world sees.

INFRASTRUCTURE AT SPEED. NEWSGUARD AI TRACKING CENTER, JUNE 2026

3,749

AI-GENERATED
CONTENT-FARM NEWS SITES

16

LANGUAGES COVERED
BY THE TRACKER

300 to 500

NEW SITES ADDED PER MONTH

Earlier counts from the same tracker stood at around a third of this total, which makes the number more useful as a rate than as a fact.

Content at volume

Hundreds of variants of a message, tailored by language, region and audience, generated in minutes.

Personas rather than bots

Accounts with consistent histories, coherent opinions and the ability to hold a conversation, replacing the crude automated accounts that were once easy to identify.

Amplification

Coordinated networks that generate not just posts but reply chains and engagement patterns that mimic genuine interaction.

Translation and reach

The same operation running simultaneously across languages that previously required separate teams.

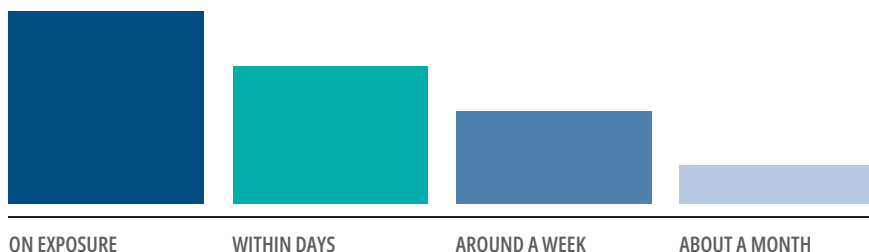
WHY DISTRIBUTION MATTERS MORE THAN CONTENT

Manipulation increasingly happens at the level of distribution rather than of the message. Low-quality content pushed hard through engineered amplification outperforms high-quality content that nobody pushes. And a message that is legitimate in itself becomes manipulation once the apparent support behind it is manufactured. This is uncomfortable for journalists because the professional reflex is to interrogate the claim. Interrogating the claim is no longer sufficient. The spread is the story as often as the content is.

AUDIENCES ARE NO LONGER ONLY HUMAN

A significant share of what you publish is now read first by machines: search crawlers, the retrieval systems behind AI assistants and the pipelines that assemble training data. Operators deliberately target open sources (Wikipedia, Reddit and social platforms among them) with false material, so that these systems absorb it and repeat it to human audiences with the authority of a neutral answer. When you publish a correction, part of the audience you are writing for is a machine that will summarise you to someone else.

PREBUNK PROTECTION DECAYS; REPETITION IS WHAT SUSTAINS IT



Indicative shape, not measured values. Text and video prebunks retain useful effect for roughly a month, and game-based ones fade faster (Capewell et al., 2024). Repetition is not optional. It makes the intervention work.

PREBUNKING AND ITS LIMITS

Prebunking is the practice of exposing an audience to a weakened form of a manipulation technique, together with a refutation, before they encounter the real thing.

It is not either/or

Prebunking and debunking do different jobs and you need both.

It decays

Protection begins to weaken within days. Text and video prebunks retain useful effect for around a month; game-based ones fade faster (Capewell et al., 2024). Repetition is not optional. It is what makes the intervention work, and its effects last longest when the reinforcement targets memory of the intervention itself (Maertens et al., 2025).

Reach governs everything

A prebunk that does not reach the same audience as the manipulation has no effect, however well designed. Section 06 applies here as much as to corrections.

A NOTE ON THE EVIDENCE BASE

The cases in this guide come largely from African elections and information environments. This is because several of these cases are among the most thoroughly documented instances of AI-assisted political manipulation anywhere. In addition, the structural conditions (mobile-first access, closed messaging dominance and thin platform moderation investment in local languages) shape both what is observable and what is fixable.

WHERE TO GO NEXT SOURCES CITED

VERIFICATION AND ARCHIVING

Bellingcat Online Investigation Toolkit — the standing reference for verification and archiving tools, updated continuously.

DeepFake-o-Meter, University at Buffalo Media Forensics Lab — free, open-source synthetic media detection.

Wayback Machine (archive.org) and archive.today.

FACT-CHECKING ORGANISATIONS

Africa Check — South Africa and continental.

Dubawa — Nigeria and West Africa.

Moxii Africa (formerly Media Monitoring Africa).

MONITORING AND RESEARCH

NewsGuard AI Tracking Center — running count of AI-generated content-farm sites.

Institute for Security Studies (ISS) Africa — election information-integrity analysis.

Global Investigative Journalism Network and Thomson Reuters Foundation. Training resources.

PUBLISHED BY



Konrad-Adenauer-Stiftung — Media Programme Sub-Saharan Africa. Companion volume to the academic study.

The companion academic volume titled *Manufactured Publics: Artificial Intelligence, Distribution and the New Telescreen* gives the full scholarly sources.

Allen, K. and le Roux, J. (2024) Under the Influence?

Online Mis/Disinformation in South Africa's May 2024 Election. Southern Africa Report 61, December. Pretoria: Institute for Security Studies.

Allen, K. and Nehring, C. (2025) AI-Generated Disinformation in Europe and Africa: Use Cases, Solutions and Transnational Learning. Johannesburg: Konrad-Adenauer-Stiftung, Media Programme Sub-Saharan Africa.

Allen, K., Wasserman, H. and Mbekela, N. (2026) Navigating Risks and Rewards: How South African Journalists Use AI in the Newsroom. Johannesburg: Konrad-Adenauer-Stiftung with the Centre for Information Integrity in Africa, Stellenbosch University.

Capewell, G., Maertens, R., Remshard, M., van der Linden, S., Compton, J., Lewandowsky, S. and Roozenbeek, J. (2024) 'Misinformation interventions decay rapidly without an immediate posttest', *Journal of Applied Social Psychology*, 54, pp. 441–454.

Konrad-Adenauer-Stiftung Uganda and South Sudan with Agora (2026) Toolkit for Countering Disinformation: Empowering Citizens in the Quest for Truth. Kampala: Konrad-Adenauer-Stiftung.

Maertens, R. et al. (2025) 'Psychological booster shots targeting memory increase long-term resistance against misinformation', *Nature Communications*, 16(2062).

NewsGuard (2026) AI Tracking Center.

Wack, M., Linvill, D. and Warren, P. (2024) Old Despots, New Tricks. Clemson, SC: Clemson University Media Forensics Hub.

Reporting on AI-driven disinformation

A working guide for journalists

**Heather A. Thuynsma, Yanga Malotana
and Christopher Nehring**

This book is the companion volume to the monograph *Manufactured Publics: Artificial Intelligence, Distribution and the New Telescreen*. This guide is for journalists working to deadline and is organised around what to do. The two volumes can be read independently.

The monograph is for readers who want the argument behind the advice. It discusses why the threat has taken the shape it has, what the evidence supports and what follows for journalism, for regulation and for the study of African information environments.

Readers who want the practical material will find it in this guide.

NON-FICTION

ISBN 978-1-0671823-4-2



9 781067 182342